An Effective Machine Learning (ML) Approach to Quality Assessment of Voice over IP (VoIP) Calls

Elena Cipressi*†, Maria Luisa Merani†‡
*Empirix, Italy †Dipartimento di Ingegneria "Enzo Ferrari",
University of Modena and Reggio Emilia, Modena, Italy
‡CNIT, Consorzio Nazionale Interuniversitario per le Telecomunicazioni, Parma, Italy
e-mail: ecipressi@empirix.com, marialuisa.merani@unimore.it

***Abstract*—This work puts forward a supervised ML technique to determine the Quality of Experience (QoE) of VoIP calls. It takes its beginning from an investigation on VQmon®, an enhanced E-model version that estimates the quality of IP-based voice calls adopting an objective approach. The current study demonstrates VQmon® shortcomings via a comparison between the Mean Opinion Score (MOS) values this technique predicts and the actual average ratings collected from a subjective listening quality campaign. It proposes to deploy Ordinal Logistic Regression (OLR) for speech quality assessment, and results disclose that OLR outperforms popular ML algorithms, in accuracy and confusion matrices.**

***Index Terms*—Machine Learning (ML); Mean Opinion Score (MOS); Quality of Experience (QoE); Voice over IP (VoIP); Speech Quality Assessment.**

I. Introduction

A. Rationale and Contribution

Quality of Experience (QoE) of VoIP calls is a relevant topic in the realm of contemporary networks, given the widespread adoption of VoIP in wired scenarios. It is even more significant in cellular networks, where VoIP counterpart, VoIP over LTE (VoLTE), combined with wideband and super-wideband codecs, plays the leading role in ensuring high quality levels for voice calls in a totally IP-based setting.

In a previous study [1], we assessed the end-to-end transmission quality of several millions of VoLTE calls employing VQmon® [2], an objective, non-intrusive tool, that enhances the standardized E-Model [3]. Tools like VQmon® are quite popular on the service assurance rim, as they can be easily integrated in proprietary software by mobile network operators. Yet, they are quickly becoming obsolete, given the complexity and heterogeneity of modern communication systems [4].

Taking the last remark as its foremost motivation, the aim of this letter is two-fold: (i) first, to quantify VQmon® limits in QoE assessment of VoIP calls that employ a wideband voice codec; (ii) then, to overcome such limits proposing the adoption of a supervised ML approach. With reference to the latter point, the current study demonstrates to what extent OLR performs better than other popular state-of-the art ML solutions. Therefore, the work proves that the OLR algorithm is well suited to model the human level of preferences expressed on an ordinal rating scale.

In order to achieve the goals stated above, a subjective listening campaign has been conducted in a controlled environment; the transmission of wideband, high quality VoIP calls has been repeatedly mimicked, collecting network metrics and several categorical features of the volunteers participating in the quality assessment test. Participants have been asked to rate the listening quality of test calls and the test outcomes have first of all disclosed VQmon® flaws. Most importantly, the test results have allowed highlighting the benefits of the proposed ML approach, which is fast like non-intrusive methods, as it automates speech quality prediction, and trustable, being built on a subjective basis that can be retrained several times upon customer availability and network adjustments. Our contributions therefore embrace the goals of mobile operators and network monitoring companies, that not only mandate for effective monitoring tools, but also for ease of deployment on millions of VoIP calls. In future networks, the OLR predictions could also help customer experience and service quality managers to identify potential network issues, on the basis of the estimated QoE values.

Finally, the study highlights that the conventional five score scale for call quality classification is often perceived as excessive by test participants. In the limiting case where ratings are collapsed on a coarse binary scale, OLR and alternative ML models are verified to guarantee a very high and comparable accuracy level.

B. Related Work

In the past, a few solutions based on advanced statistics and ML models such as Bayesian Classifier [5], Artificial Neural Networks [6] and Random Neural Networks [7] have been proposed to predict VoIP speech quality. As a recent example belonging to this category, the study in [8] compares the performance of different ML classifiers, considering packet loss, narrow-band codec type, language and gender as features. All the previously cited works assume as learning basis (equivalently termed ground-truth) the quality ratings that the Perceptual Evaluation of Speech Quality (PESQ) technique [9] provides. PESQ is an algorithm for narrow-band voice evaluation; it is objective, i.e., it automatically evaluates speech quality with no involvement of human subjects, and it is double-reference, as it compares the received voice signal against the clean, original signal. However, one relevant drawback inherent to the choice of employing PESQ outcomes as ground truth is that the estimate error affecting the reference

TABLE I
MOST FREQUENTLY EMPLOYED ACRONYMS AND THEIR MEANING

Acronym	Term in full
DT	Decision Tree
LinReg	Linear Regression
LogReg	Logistic Regression
ML	Machine Learning
MLR	Multinomial Logistic Regression
OLR	Ordinal Logistic Regression
RF	Random Forest

technique propagates to the learning algorithm. Alternative studies, like [10], considered as ground-truth the subjective Mean Opinion Score (MOS), that the ITU defines as the arithmetic mean of a collection of single user opinion scores [11]. Yet, the arithmetic mean might represent a rough approximation when judging the quality of VoIP calls: it inevitably smooths out the quality score that a specific user assigns the call under certain network conditions. Lastly, P. Charonyktakis et al. [12] designed a modular algorithm that uses multiple ML models, including Decision Trees and Support Vector Regression, and relies on an optimized technique, termed nested cross validation, to select the best classifier. This study adopts both subjective tests and PESQ to rate the actual QoE of narrow-band VoIP calls.

Partly in analogy to the contribution in [12], the present letter concentrates on the subjective experience of single users as ground truth. Differently from [12] and previous works, our study proposes to handle the rating of the call quality experienced by the single user as an intermediate problem between regression and classification. It therefore suggests to exploit a specific algorithm, the so-called Ordinal Logistic Regression (OLR), and it benchmarks its performance against some of the most popular ML methods already utilized in the works cited above, highlighting its better accuracy. Further, our contribution concentrates on wide-band, high definition voice, which is of paramount importance in VoLTE, as well as in 5G networks. To the best of our knowledge, all the investigations on VoIP QoE presented so far in literature are centered on the adoption of narrow-band codecs, that work on audio frequencies in the 300-3400 Hz range. However, all modern applications relying on telephony audio employ wideband and super-wideband codecs, which extend the maximum operating frequency to 7 and 22 KHz, respectively. We therefore choose to concentrate on wideband codecs.

II. BACKGROUND AND SETTING

This section is intended to build a concise background on the ML models employed in this investigation, and to provide an overview of the experiment setting and design. An extensive explanation of the selected algorithms and of their implementation details can be found in [13] and [14]. The list of the acronyms most frequently employed in this letter is reported in Table I.

A. Prediction Models

The distinctive characteristic of supervised learning is that the target label to predict is known (e.g., in this work we know the QoE labels), and this information is explicitly used in the learning process. Supervised learning approaches can be further distinguished in classification and regression solutions. We refer to classification when the target variable is a class, as in the examined problem. In particular, the rating assigned to call quality is organized on an ordered scale featuring five score values (classes): 1 (bad), 2 (poor), 3 (fair), 4 (good) and 5 (excellent). Among ML classification algorithms, the Decision Tree classifier (DT) is a solution that produces interpretable models and it is widely employed for this distinctive feature: its goal is to create a model that learns from simple if/else rules inferred from data. To build a tree, the algorithm searches over all possible paths and finds the one that is the most informative about the target variable. An enhancement to DT is Random Forest (RF), fitting a number of DT classifiers on various subsets of the dataset. RF relies upon an ensemble of trees to improve predictive accuracy. Trees are easily visualized and interpreted, but their main drawback is that they neglect any ordered trait of the target feature.

Differently from classification, regression predicts a continuous outcome. The reference model is Linear Regression (LinReg), utilized to find the relation between two or more continuous variables. Logistic Regression (LogReg) replaces LinReg when the target is no longer continuous and is expressed as a dichotomous variable. Its generalization to more than two classes is Multinomial Logistic Regression (MLR).

Lastly, OLR represents an intermediate approach between classification and regression, and it is our belief that it can successfully fit the present problem of predicting *ordered* classes of QoE. As a matter of fact, OLR handles labels that are both discrete as in classification, and ordered as in linear regression. Its complex mathematical formulation is based on the generalized linear model (GLM), well-detailed in [15] and [16].

B. Experiment Setting and Design

Fig.1 portrays the end-to-end setting of our experiment. Calls were generated by Hammer®, a proprietary platform by Empirix [17] that emulates software agents initiating and accepting VoIP calls and establishing an SIP/RTP session for every call. One Hammer was installed on the Virtual Machine (VM) of a Windows PC, acting as the caller (Hammer A), a second Hammer was installed on the VM of a second PC, representing the callee; a Linux-based, Ubuntu VM on a third PC routed packets from the caller to the callee, and also acted as a source of impairments through Netem [18], a network simulator available in Linux kernels. All PCs belonged to the same Gigabit Ethernet LAN.

We chose to deliver the short audio stream "You will have to be very quiet", encoded through Adaptive Multi Rate WideBand (AMR-WB) [19] (mode 25.85 kb/s) and fully compliant with ITU-T guidelines about subjective listening tests [11]. Each call featured the same audio stream. Through Netem [18], we intervened on the one-way delay and packet loss to simulate the typical impairments of real networks. In detail, given the ITU-T G.1010 document [20], that suggests

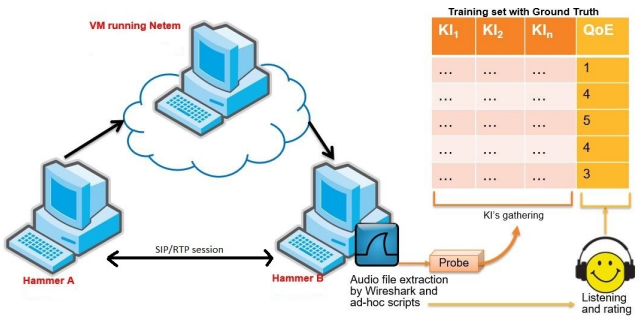


Fig. 1. End-to-end experiment workflow.

the tolerated values of one-way delay (lower than 400 ms) and packet loss rate (lower than 1%) for conversational audio, we combined four profiles of packet loss (random, uniformly distributed losses with rates 0%, 0.5%, 1% and 2%) with three profiles of one-way delay (Gaussian distributed with mean and standard deviation equal to (0 ± 0)ms, (150 ± 25)ms and (400 ± 25)ms), thus obtaining twelve scenarios. At the callee side, the jitter buffer was instantiated to receive packets with a fixed inter-packet delay. The received files were collected in a Wireshark [21] compatible format, and sent to a proprietary probe, where they were processed and then exported. Since we operated in a virtual environment, we made use of *ad-hoc* scripts[1] to extract the audio trace in a listenable format.

We next conducted a subjective listening campaign, and designed the listening experiments in accordance to ITU-T guidelines [11]. Among the available quality assessment methods, we adopted the popular Absolute Category Rating (ACR) test, because of its reliability and fast implementation. In ACR subjective tests, users are asked to evaluate calls, presented only once, and have to rate the listening quality, i.e., their QoE, on the ordered scale with 5 possible score values. For the experiment, a pool of 56 participants was recruited on a voluntary basis in the first half of 2019. Every listener was asked to evaluate the quality of 12 calls, corresponding to the received audio streams in the 12 scenarios described above. Before starting the survey, we additionally asked volunteers to answer a few questions, namely, to indicate their gender, age and the type of headset employed during the test. These categorical features uniquely characterized each participant, along with the rating she/he attributed to the quality of the calls. In addition, we encouraged users to share their feedback.

At the end of the experiment, we collected a total of 672 evaluations. Because of the arbitrary property of subjective tests, it is known that some ratings might have been assigned in an inappropriate manner. Thus, we grouped call scores by call identifier and applied the popular DBscan algorithm [22] to detect outliers among the evaluations collected for each call. DBscan found a total of 55 outliers, that were removed from the initial set.

[1]https://github.com/Spinlogic/AMR-WB_extractor

III. Experimental Results

A. Data Set Preprocessing

Network side, the input features we gathered from the testbed in Fig.1 included the actual network metrics associated with each evaluated call, that is, the following numerical features: average and maximum jitter, number of received packets, packet loss rate, out-of-sequence packets and duplicated packets. The input features coming from the test participants were the categorical features, that is, their age, gender and type of headset. Most importantly, we collected the rating each participant attributed to the quality of the calls, i.e., their QoE scores.

To minimize the risk of injecting noise into the model, we firstly determined the most informative features with respect to the target label, i.e., the QoE score. This also serves as a preliminary action to understand the impact they have on QoE. According to Pearson's correlation test [23], numerical features exhibiting a p value greater than 0.01 were considered insignificant. We therefore neglected the number of received packets and the number of duplicated packets. Given the relatively modest number of examined settings for the test, we additionally "flagged" the packet loss rate as a binary variable: we stated that it was *present* in any scenario where it took on values greater than 10^{-2}, otherwise it was interpreted as *absent* (for the considered scenarios, this corresponds to values lower than 10^{-3}). Lastly, as the numerical features span quite different ranges, we rescaled them, in order not to privilege one over others (e.g., maximum jitter over out-of-sequence packets). Because of the relatively low number of research participants, we decided to include all the categorical features in the present study, as it is not possible to firmly state that QoE is independent of them.

B. Exploratory Investigation and Performance Assessment

Preliminarily, we investigated the reliability of VQmon® when assessing the quality of wide-band VoIP calls; for doing so, we compared the MOS values that VQmon® provides against those determined from the actual subjective ratings; adhering to MOS definition, we computed the latter value as the average of the individual ratings that different users assigned to the same call. Fig. 2 shows how far VQmon® MOS values (red squared markers in the figure) are from their experimental counterparts (blue circles) for the 12 synthetic calls evaluated by the users. Vertical bars refer to the standard

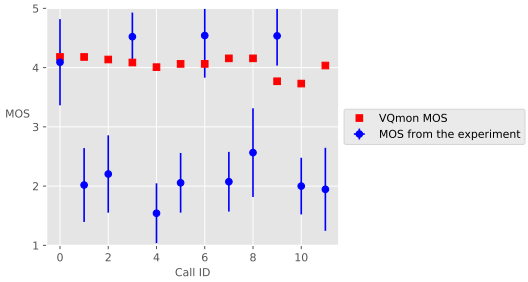


Fig. 2. VQmon® MOS and Users MOS per Call ID.

TABLE II
OLR AND DT CONFUSION MATRICES

		Predicted by OLR				
		1	**2**	**3**	**4**	**5**
Observed	**1**	2	15	0	0	0
	2	2	39	3	0	0
	3	0	13	6	2	3
	4	0	0	0	12	6
	5	0	0	0	3	17

		Predicted by DT				
		1	**2**	**3**	**4**	**5**
Observed	**1**	0	17	0	0	0
	2	0	44	0	0	0
	3	0	19	0	5	0
	4	0	0	0	4	14
	5	0	0	0	5	15

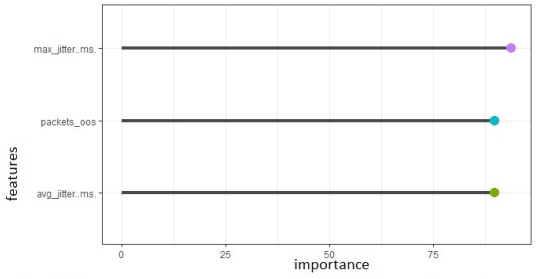


Fig. 3. Feature importance for DT (%).

deviation of the users MOS. The results reported in this figure clearly demonstrate that VQmon® cannot predict the actual call quality, and further motivates us to explore the effectiveness of a user-driven methodology that exploits ML tools.

Given the presence of ordered classes, i.e., the five possible QoE scores, we deliberately focused on OLR as a promising candidate among the alternative ML algorithms. To validate the goodness of such a choice, we considered a random split of the QoE scores, employing 80% of them as the training set and the remaining 20% as the test set and first benchmarked OLR classification accuracy against that of the Random Classifier (RC), DT, RF and MLR. We decided to exclude Neural Networks from our investigation because of the relatively few training data available [24]. Moreover, we did not consider Support Vector Machines (SVM) models either, because they do not perform well with unbalanced classes [25], as is the case here.

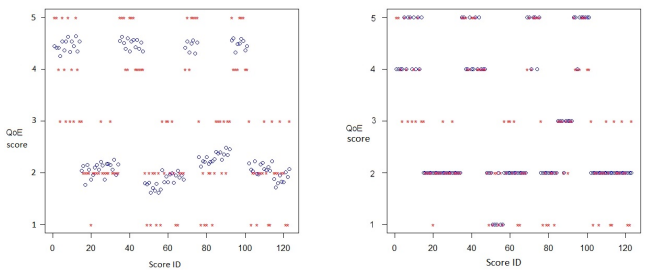
Recalling that accuracy is defined as the percentage of correct predictions to the total number of test samples, we found that OLR outperforms all other algorithms. Its accuracy is 61%, almost four times the RC accuracy, which is only 16%. The OLR accuracy is higher by ten percentage points than DT (51%) and higher by 9% than MLR (52%), where we emphasize that the latter two algorithms do not take into account class ordering.

The confusion matrix generalizes the concept of accuracy: every row represents the instances in an actual class and every column represents the instances in a predicted class, so that the ideal confusion matrix has zero elements everywhere except for the main diagonal, meaning that all the predicted instances coincide with the actual observations. Table II compares OLR and DT confusion matrices, revealing that OLR better captures intermediate opinions (3 and 4 QoE values), that are more likely related to each test participant and her/his set of unmeasurable characteristics (e.g., mood, tolerance level), whereas DT limits its prediction to three out of five QoE classes (2, 4 and 5). We decided not to include the detailed results of RF in our analysis, as we verified that RF performs worse than DT, its accuracy being 48%. This is explained by the limited size of the data set (617).

For DT, Fig. 3 shows the importance of the three most relevant features in our dataset, namely, maximum jitter, out-of-sequence packets and average jitter. We derived this indication for a classification tree algorithm that employs Gini impurity [26] to rule the split of the data samples down the tree. Although not reported, the features statistically having the major impact on QoE are the same when the OLR approach is taken. In this regard, note that a supervised ML algorithm builds the model by tuning its parameters on the training data; these parameters weight the importance different features have and how they are combined together in the model. For DT (and RF as well), understanding the features that most affect the target variable is inherently straightforward, due to the way the tree is built. On the other hand, OLR fits both a coefficient vector and a set of thresholds to the training dataset [15], and therefore the feature interpretation is not as immediate.

To exclude that this study had to be approached as a linear problem, we further considered LinReg as an alternative baseline. We therefore extended the domain of the target label QoE from integer to real, thus removing the concept of classes. The subjective QoE scores (red squares) and the predicted values (blue circles) are reported in Fig. 4(a) for LinReg and in Fig.4(b) for OLR. They allow to compare the performance of LinReg and OLR, disclosing that LinReg is unable to predict intermediate results, whereas OLR can.

Lastly, it is interesting to outline that out of 56 research participants, almost half of them pointed out that five classes were too many to evaluate the QoE of the test calls, which might more simply be rated as poor or good. Adhering to this rationale, we remapped the five original classes into two, class 0 collecting the previous 1 and 2 classes, and class 1, merging classes 3, 4 and 5, so as to reduce the problem to binary classification. As such, the concept of ordering no longer holds, and the binary counterpart of OLR is LogReg. When taking this approach, both accuracy and confusion matrix remarkably improve. As is to be expected, LogReg and DT



(a) LinReg (b) OLR

Fig. 4. LinReg and OLR performance.

TABLE III
DT AND LOGREG CONFUSION MATRICES (BINARY CLASSIFICATION)

	Predicted (DT)		
		0	1
Observed	0	62	0
	1	21	41

	Predicted(LogReg)		
		0	1
Observed	0	60	2
	1	18	44

exhibit similar performance. In detail, LogReg accuracy stands at 83% and by inspecting LogReg and DT confusion matrices reported in Table III, we observe the prevalence of correctly predicted instances. Although not reported on the table, we also tested the RF model for this reduced binary classification problem. We coherently obtained the same accuracy and confusion matrix as DT. This is not surprising, and it can be explained by the reduction in the problem cardinality and complexity. To complete our proposal, we devote a few words to the network location where a monitoring tool based on our approach could be deployed. With reference to LTE, the M_b standard interface at the border between the Packet data network GateWay (PGW) in the Evolved Packet Core (EPC) and the IP Multimedia System (IMS) is the most proper choice. This is the interface where probes are conventionally placed to monitor VoLTE traffic, capturing relevant network parameters for each voice flow. Once properly tuned, the new tool would allow associating a QoE level estimate to every flow in nearly real-time. Then, the automatic inspection of voice traces and of the corresponding QoE values would serve to identify, e.g., critical areas where the QoE guaranteed to network subscribers is not adequate.

IV. CONCLUSIONS

This work has conducted a subjective campaign of quality assessment on artificially generated VoIP calls, collecting the values of network metrics associated with each test call, some categorical features of the participants and their QoE scores. The shortcomings of a conventional objective, no-reference model when assessing speech quality of VoIP wide-band calls was first demonstrated. It was next proposed to adopt a customer-driven, ML approach to correlate network-oriented features and human-related aspects to the levels of QoE that listeners perceive. OLR has been proved to be the best algorithm to model the examined problem. It guarantees a high prediction accuracy, owing to its ability to capture the ordinal behavior of subjective experience. The study has additionally provided an insight into the difficulties of utilizing a five level scale to evaluate VoIP QoE, often perceived and described by test participants as poor or good. When handling the quality assessment problem as binary instead of ordinal, it has been shown that LogReg, the binary counterpart of OLR, just like different algorithms (e.g., RF) guarantee reliable and similar predictions.

REFERENCES

[1] E. Cipressi and M. L. Merani, "A Comparative Study on the Quality of Narrow-Band and Wide-Band AMR VoLTE Calls," in *IEEE 15th International Wireless Communications Mobile Computing Conference (IWCMC)*, June 2019, pp. 1273–1278.

[2] Telchemy, "VQmon," 2019. [Online]. Available: https://www.telchemy.com/vqmon.php

[3] ITU-T, "The e-model: a computational model for use in transmission planning," *ITU-T Recommendation G.107*, 2015.

[4] S. Jelassi, G. Rubino, H. Melvin, H. Youssef, and G. Pujolle, "Quality of Experience of VoIP Service: A Survey of Assessment Approaches and Open Issues," *IEEE Communications Surveys Tutorials*, vol. 14, no. 2, pp. 491–513, Second 2012.

[5] F. Rahdari and M. Eftekhari, "Using Bayesian classifiers for estimating quality of VoIP," in *16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012)*, May 2012, pp. 348–353.

[6] L.Sun and E. C. Ifeachor, "Perceived speech quality prediction for voice over IP-based networks," in *2002 IEEE International Conference on Communications. Conference Proceedings. ICC 2002*, vol. 4, April 2002, pp. 2573–2577.

[7] W. Cherif, A. Ksentini, D. Negru, and M. Sidibe, "A_PSQA: PESQ-like non-intrusive tool for QoE prediction in VoIP services," in *2012 IEEE International Conference on Communications (ICC)*, June 2012, pp. 2124–2128.

[8] R. S. Alkhawaldeh, S. Khawaldeh, U. Pervaiz, M. Alawida, and H. Alkhawaldeh, "NIML: non-intrusive machine learning-based speech quality prediction on VoIP networks," *IET Communications*, vol. 13, no. 16, pp. 2609–2616, 2019.

[9] ITU-T, "P.862 : Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation P.862*, Approved in February 2001.

[10] S. Mohamed, F. Cervantes-Perez, and H. Afifi, "Integrating networks measurements and speech quality subjective scores for control purposes," in *Proceedings IEEE INFOCOM 2001. Conference on Computer Communications. Twentieth Annual Joint Conference of the IEEE Computer and Communications Society (Cat. No.01CH37213)*, vol. 2, April 2001, pp. 641–649.

[11] ITU-T, "Methods for subjective determination of transmission quality," *ITU-T Recommendation P.800*, 1996.

[12] P. Charonyktakis, M. Plakia, I. Tsamardinos, and M. Papadopouli, "On User-Centric Modular QoE Prediction for VoIP Based on Machine-Learning Algorithms," *IEEE Transactions on Mobile Computing*, vol. 15, no. 6, pp. 1443–1456, June 2016.

[13] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.

[14] A. Agresti, *Categorical data analysis*, ser. A Wiley-Interscience publication. New York [u.a.]: Wiley, 1990.

[15] C. Winship and R. D. Mare, "Regression models with ordinal variables," *American Sociological Review*, 1984. [Online]. Available: http://www.asanet.org/journals/asr/

[16] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, "Ordinal regression methods: Survey and experimental study," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 127–146, Jan 2016.

[17] "Hammer." [Online]. Available: https://www.empirix.com/products/hammer/

[18] "Netem." [Online]. Available: https://wiki.linuxfoundation.org/networking/netem

[19] GSMA, "Adaptive multirate wide band," *GSMA Official Document IR.36*, 21 February 2013.

[20] ITU-T, "G.1010 : End-user multimedia QoS categories," *ITU-T Recommendation G.1010*, Approved in November 2001.

[21] "Wireshark." [Online]. Available: https://www.wireshark.org/

[22] M. Hahsler, "Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms - R package," 2019. [Online]. Available: https://www.rdocumentation.org/packages/dbscan/versions/1.1-4

[23] W. Kirch, Ed., *Pearson's Correlation Coefficient*. Springer Netherlands, 2008. [Online]. Available: https://doi.org/10.1007/978-1-4020-5614-7_2569

[24] A. Ng, *Machine Learning Yearning*. Online Draft, 2017. [Online]. Available: http://www.mlyearning.org/,/bib/ng/ng2017mlyearning/Ng_MLY01_13.pdf

[25] V. Palade, *Class imbalance learning methods for support vector machines*. Wiley Online Library, 2013.

[26] R. Documentation, "Caret v6.0-85," 2019. [Online]. Available: https://www.rdocumentation.org/packages/caret/versions/6.0-85/topics/varImp